

Stochastic resetting in backtrack recovery by RNA polymerases

Édgar Roldán^{*,1,2,3} Ana Lisica,^{4,5} Daniel Sánchez-Taltavull,⁶ and Stephan W. Grill^{*1,4,5}

¹Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Str. 38, 01187 Dresden, Germany.

²Center for Advancing Electronics Dresden, cfaed, Dresden, Germany.

³GISC - Grupo Interdisciplinar de Sistemas Complejos, Madrid, Spain.

⁴BIOTEC, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany.

⁵Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstraße 108, 01307 Dresden, Germany.

⁶Regenerative Medicine Program, Ottawa Hospital Research Institute, Ottawa, K1H 8L6, Canada.

Transcription is a key process in gene expression, in which RNA polymerases produce a complementary RNA copy from a DNA template. RNA polymerization is frequently interrupted by backtracking, a process in which polymerases perform a random walk along the DNA template. Recovery of polymerases from the transcriptionally-inactive backtracked state is determined by a kinetic competition between 1D diffusion and RNA cleavage. Here we describe backtrack recovery as a continuous-time random walk, where the time for a polymerase to recover from a backtrack of a given depth is described as a first-passage time of a random walker to reach an absorbing state. We represent RNA cleavage as a stochastic resetting process, and derive exact expressions for the recovery time distributions and mean recovery times from a given initial backtrack depth for both continuous and discrete-lattice descriptions of the random walk. We show that recovery time statistics do not depend on the discreteness of the DNA lattice when the rate of 1D diffusion is large compared to the rate of cleavage.

PACS numbers: 05.40.-a, 87.10.Mn

I. INTRODUCTION

Transcription of genetic information from DNA into RNA is the first step of gene expression and is fundamental for cellular regulation. The process is performed by macromolecular enzymes called RNA polymerases that move stepwise along a DNA template and produce a complementary RNA (see Fig. 1). Transcription elongation is often interrupted by pausing and backtracking, a reverse movement of the RNA polymerase on the DNA template that displaces the RNA 3' end from the active site and leaves the enzyme transcriptionally inactive [1–5]. The polymerase recovers from a backtrack when it realigns the 3' end of the RNA with its active site.

In a backtrack, polymerases perform random walk on the DNA template [6]. The recovery of the polymerase from a backtracked state, i.e. *backtrack recovery*, results from the kinetic competition between two mechanisms [7]: polymerases can either recover by performing a random walk along the DNA until it returns to the elongation competent state [6, 8–12] or by cleavage of the backtracked RNA which generates a new RNA 3' end in the active site [13–15]. The cleavage reaction can be performed by intrinsic cleavage mechanisms or it can be assisted by a transcription factor, TFIIS [16–18].

The stochastic motion of the RNA polymerase in a backtrack was previously measured with single-molecule

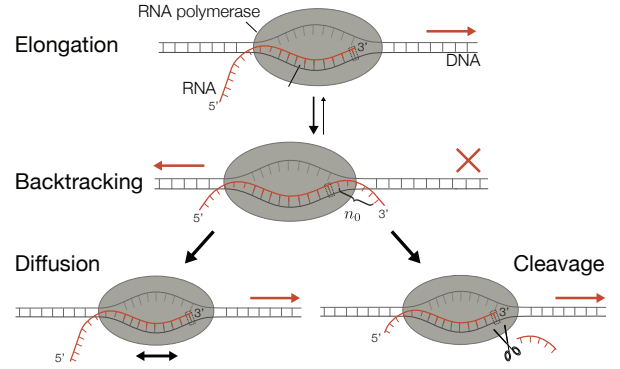


FIG. 1: Scheme of transcription elongation (top), backtracking (middle) and backtrack recovery (bottom) processes. The 3'-end of the RNA is aligned with the active site of the polymerase (dashed square) during elongation, but displaced in the backtracked state. This makes the polymerase transcriptionally inactive, as it can not add new nucleotides to the 3'-end of the RNA. Here, n_0 represents the number of backtracked nucleotides. Backtrack recovery can proceed through 1D diffusion of the polymerase along the DNA template, until the 3-end of the RNA realigns with the active site or through cleaving of the backtracked RNA and production of a new 3'-end that is aligned with the active site.

optical tweezers and described using continuous time Markov processes [6, 8, 10–12, 19]. Specifically, backtracking has been treated as a diffusion process in continuous space [6] but also as a hopping process over a discrete lattice of nucleotides [9–12, 20–22]. However, it remains unclear which aspects of the backtracking process depend on the discreteness of the position lattice and

*Correspondence should be addressed to Édgar Roldán (Email: edgar@pks.mpg.de), and Stephan W. Grill (Email: stephan.grill@biotec.tu-dresden.de).

which can be described with a diffusion process.

Here we present both discrete and continuous-space descriptions of backtrack recovery and investigate to which extent a diffusion process is a good approximation of the polymerase dynamics during a backtrack. We present a solvable stochastic model of RNA polymerase backtrack recovery that includes both diffusion and cleavage and study its main statistical features. The process shares similarities with recent development on diffusion processes with *stochastic resetting* introduced in Ref. [23]. In such problems a particle undergoes Brownian diffusion but can also stochastically reset its position [23–30]. The mean first-passage time to an absorber can be determined analytically, which depends on the statistics of resetting [23, 24, 27, 28]. A backtracked RNA polymerase undergoes a random walk to the elongation competent state while also resetting its position via cleavage, and we here determine the first-passage time properties of this variant of a ‘diffusion with stochastic resetting’ process.

In experiments, deep backtracks are readily identified, and it is possible to determine accurately the time it takes the polymerase to recover from a backtrack of a certain depth [7]. Therefore, we here determine the recovery time τ_{rec} , defined as the first-passage time of a random walker to reach an absorbing barrier with the walker starting at a given ‘initial’ backtracking depth. We derive exact expressions of relevant statistics, such as the mean time to recover from a backtrack, or mean recovery time, for both continuous and discrete stochastic models. Remarkably, we find that for the case when the hopping rate is much larger than the cleavage rate both discrete and continuous descriptions can be used concurrently to describe the statistics of backtrack recovery from short and long initial backtrack depths.

II. DISCRETE MODEL: HOPPING PROCESS WITH CLEAVAGE

We first describe the recovery of an RNA polymerase from a backtrack as a continuous-time 1D hopping process on a semi-infinite discrete lattice. Each state of the lattice $n \in [1, 2, 3, \dots, \infty)$ represents the number of nucleotides backtracked by the polymerase (see Fig. 3). For example, $n(t) = 3$ means that at time t the polymerase has backtracked 3 nucleotides. In our model, polymerases can jump between adjacent states with hopping rate k and can cleave an arbitrarily long RNA transcript with a cleavage rate k_c . We consider that no external forces bias the hopping rates of the polymerase on the lattice. Cleavage is represented by an instantaneous jump or stochastic reset [23–30] to the elongation-competent state located in $n = 0$. The elongation-competent state is considered as an absorbing state because the probability to backtrack after cleavage is very low [11]. Our discrete model is a variant of the hopping models introduced by Depken *et al.* in Refs. [9, 32].

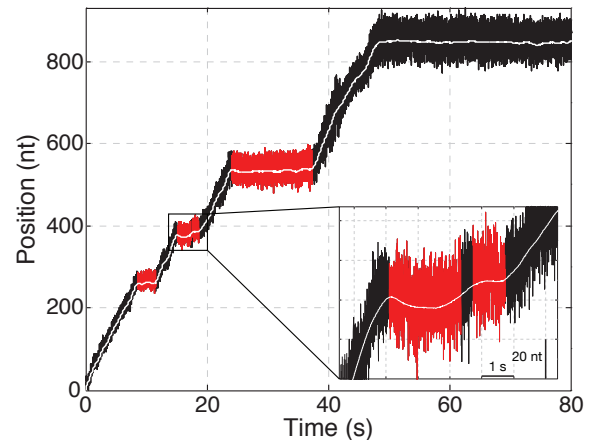


FIG. 2: **Experimental traces of transcription elongation by RNA polymerase I.** Position of the polymerase on a DNA template (in nucleotides, nt) as a function of time obtained in a single-molecule experiment. The regions highlighted in red correspond to pauses and the inset shows a zoomed view of one of the pauses, which includes a backtrack. The original data (black) is obtained with the experimental setup described in [7, 31], using a sampling rate of 1 kHz.

The time evolution of the position of the polymerase can be described in a Master equation formalism [33]. The probability of the polymerase to be at state n at time $t \geq 0$ is given by $p_n(t)$. We consider the initial condition $p_n(0) = \delta_{n,n_0}$, that is, polymerases are initially positioned at $n_0 \geq 1$. The dynamics of the probability of the polymerase to be at a given state at time t is described by the following Master equation:

$$\frac{dp_1(t)}{dt} = k p_2(t) - (2k + k_c) p_1(t) \quad , \quad (1)$$

$$\frac{dp_n(t)}{dt} = k p_{n+1}(t) - (2k + k_c) p_n(t) + k p_{n-1}(t) \quad , \quad (2)$$

where $n \geq 2$. The elongation state $n = 0$ is an absorber. Recent experiments showed that the elongation rate from $n = 0$ is more than 10 times faster than the rate of backtracking by one nucleotide [11], hence we neglect the possibility to make a jump from $n = 0$ to $n = 1$. Hence, the recovery time is the first-passage time of the polymerase to reach the absorber located in $n = 0$.

Equations (1-2) can be solved exactly (see Appendix A). Using the exact solution of the model we now derive exact expressions for the recovery time distribution and the mean recovery time of a polymerase from a given initial backtracked state.

A. Recovery time distribution

Next, we derive an analytical expression for the recovery time distribution from an initial backtrack depth, n_0 .

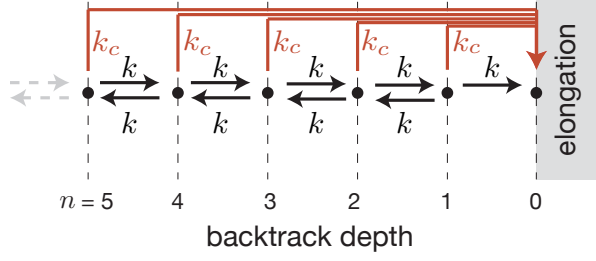


FIG. 3: **Stochastic model of backtrack recovery: hopping process with cleavage.** Each state n represents the number of backtracked nucleotides. The stochastic motion of the polymerases in a backtrack is described as a continuous-time hopping process between adjacent states with hopping rate k . Cleavage is represented as a stochastic reset to the elongation state with rate k_c . The recovery time from an initial backtrack depth n_0 is given by the first-passage time to the absorbing elongation state, $n = 0$.

We introduce a generating function

$$G(t, z) \equiv \sum_{n=1}^{\infty} p_n(t) z^{n-1} . \quad (3)$$

For $z = 0$, the generating function gives the probability to be in $n = 1$ at time t , $G(t, 0) = p_1(t)$. For $z = 1$, the generating function equals the survival probability $S(t; n_0)$ at time t starting from n_0 , $G(t, 1) = \sum_{n=1}^{\infty} p_n(t) = S(t; n_0)$.

Using the generating function, the full set of Master equations [Eqs. (1-2)] can be rewritten as a single ordinary differential equation for the generating function

$$\frac{\partial G(t, z)}{\partial t} = \left[kz - (2k + k_c) + \frac{k}{z} \right] G(t, z) - \frac{k}{z} G(t, 0) . \quad (4)$$

The initial condition $p_n(0) = \delta_{n, n_0}$ can be expressed in terms of the generating function as $G(0, z) = \sum_{n=1}^{\infty} p_n(0) z^{n-1} = \sum_{n=1}^{\infty} \delta_{n, n_0} z^{n-1} = z^{n_0-1}$. The solution of Eq. (4) with this initial condition is

$$\begin{aligned} G(t, z; n_0) &= \exp \left[\left(kz - (2k + k_c) + \frac{k}{z} \right) t \right] \\ &\times \left[z^{n_0-1} - \frac{k}{z} \int_0^t e^{-(kz - (2k + k_c) + k/z)s} G(s, 0) ds \right] . \end{aligned} \quad (5)$$

We next define $\Phi(\tau_{\text{rec}}; n_0) d\tau_{\text{rec}}$ as the probability of a polymerase to recover from an initial backtracked position n_0 in the time interval $[\tau_{\text{rec}}, \tau_{\text{rec}} + d\tau_{\text{rec}}]$. To calculate $\Phi(\tau_{\text{rec}}; n_0)$, we use the fact that a polymerase can exit a backtrack by hopping (from state $n = 1$ with rate k) or by cleavage (from any state with rate k_c). The probability density of the polymerase to reach the absorbing state at time τ_{rec} is then given by

$$\Phi(\tau_{\text{rec}}; n_0) = k G(\tau_{\text{rec}}, 0; n_0) + k_c G(\tau_{\text{rec}}, 1; n_0) . \quad (6)$$

The probability to be at the state 1 in τ_{rec} , $G(\tau_{\text{rec}}, 0; n_0)$, equals to (see Appendix A)

$$G(\tau_{\text{rec}}, 0; n_0) = e^{-(2k+k_c)\tau_{\text{rec}}} \frac{n_0 I_{n_0}(2k\tau_{\text{rec}})}{k\tau_{\text{rec}}} , \quad (7)$$

where I_{n_0} is the n_0 -th order modified Bessel function of the first kind [34].

The survival probability in τ_{rec} , $S(\tau_{\text{rec}}; n_0) = G(\tau_{\text{rec}}, 1; n_0)$ is given by

$$G(\tau_{\text{rec}}, 1; n_0) = e^{-k_c\tau_{\text{rec}}} \left[1 - k \int_0^{\tau_{\text{rec}}} e^{-2ks} \frac{n_0 I_{n_0}(2ks)}{ks} ds \right] , \quad (8)$$

which yields

$$G(\tau_{\text{rec}}, 1; n_0) = e^{-k_c\tau_{\text{rec}}} \left[1 - \frac{(k\tau_{\text{rec}})^{n_0}}{n_0 \Gamma(n_0)} H(\tau_{\text{rec}}; n_0) \right] , \quad (9)$$

where Γ is the Gamma function and H equals to

$$\begin{aligned} H(\tau_{\text{rec}}; n_0) &= {}_2F_2 \left[\left\{ n_0, n_0 + \frac{1}{2} \right\}; \{ n_0 + 1, 2n_0 + 1 \}; -4k\tau_{\text{rec}} \right] , \end{aligned} \quad (10)$$

where ${}_2F_2$ is a generalized hypergeometric function (see Ref. [34]). The recovery time distribution is obtained by substituting (7) and (9) in (6), and given by

$$\begin{aligned} \Phi(\tau_{\text{rec}}; n_0) &= e^{-(2k+k_c)\tau_{\text{rec}}} \frac{n_0 I_{n_0}(2k\tau_{\text{rec}})}{\tau_{\text{rec}}} \\ &+ k_c e^{-k_c\tau_{\text{rec}}} \left[1 - \frac{(k\tau_{\text{rec}})^{n_0}}{n_0 \Gamma(n_0)} H(\tau_{\text{rec}}; n_0) \right] . \end{aligned} \quad (11)$$

In the absence of cleavage, $k_c = 0$, the recovery time distribution becomes

$$\Phi(\tau_{\text{rec}}; n_0) = e^{-2kt} \frac{n_0 I_{n_0}(2kt)}{t} . \quad (12)$$

Note that if the polymerase is initially at $n_0 = 1$, we obtain the probability density for a pause of duration τ_{rec} ,

$$\Psi(\tau_{\text{rec}}) = e^{-2k\tau_{\text{rec}}} \frac{I_1(2k\tau_{\text{rec}})}{\tau_{\text{rec}}} , \quad (13)$$

in agreement with Depken *et al.* [9].

To verify our model, we perform numerical simulations of the hopping process with cleavage using Gillespie algorithm [35] (Fig. 4A, B). From our simulations, we calculate first-passage time distributions to the elongation state and compare it with the recovery time distribution derived in Eq. (11) (Fig. 4C, D). In the presence of cleavage, recovery can happen from arbitrarily deep backtracks. Cleavage prevents backtracks of large duration as shown by the sharp cutoff of the first-passage time distribution at large times (Fig. 4C, inset). In the absence of cleavage, deep backtracks are recovered at very

Discrete hopping model

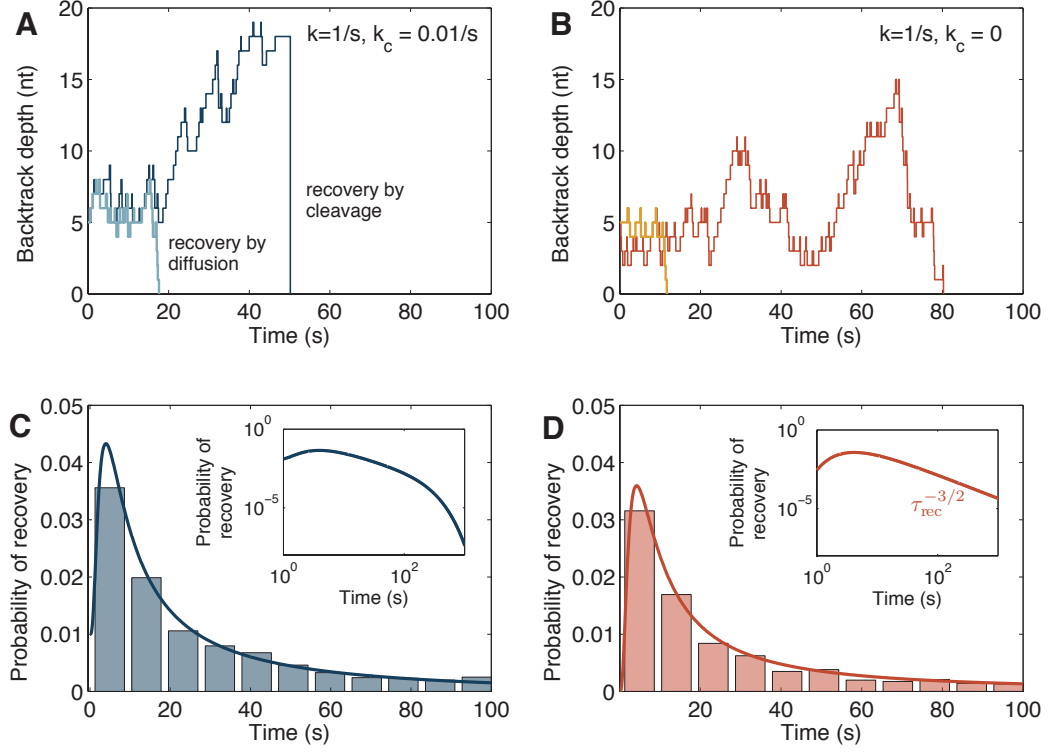


FIG. 4: Stochastic trajectories of the discrete hopping model and recovery time distributions. **A)** Sample trajectories of the hopping model with diffusion and cleavage ($k = 1/s$, $k_c = 0.01/s$) simulated using the Gillespie algorithm. The light blue trajectory represents a polymerase that recovers by diffusion, and the dark blue trajectory a polymerase that recovers by cleavage. **B)** Sample trajectories for the discrete model with only diffusion, $k = 1/s$, $k_c = 0$, obtained using the Gillespie algorithm. **C)** Recovery time probability density for the case where $k = 1/s$ and $k_c = 0.01/s$. The bars are obtained from histograms of 1000 numerical simulations and the curve is the exact expression given by Eq. (11). The inset shows a log-log plot of the recovery time distribution for long recovery times. **D)** Numerical and analytical probability density of the recovery time for the case where, $k = 1/s$, $k_c = 0$. The inset shows the tail $\tau_{\text{rec}}^{-3/2}$ of the distribution at long times. In all cases the initial backtrack depth was set to $n_0 = 5$.

large times, with a power-law tail $\Phi(\tau_{\text{rec}}; n_0) \sim \tau_{\text{rec}}^{-3/2}$ (Fig. 4D, inset). The probability density function obtained from numerical simulations in both cleavage assisted (Fig. 4C) and cleavage deficient case (Fig. 4D) agree with the theoretical expression of the recovery time distribution derived here in Eq. (11).

B. Mean recovery time

The mean recovery time $\langle \tau_{\text{rec}} \rangle$ is a useful statistic that can be measured experimentally in single-molecule experiments. Moreover, the mean recovery time can provide a quantitative measure of kinetic rates of backtrack recovery, as shown in Ref. [7]. The mean recovery time can be

obtained from Eq. (11), and equals to

$$\langle \tau_{\text{rec}} \rangle = \frac{1}{k_c} \left[1 - \left(\frac{\sqrt{\frac{4k}{k_c} + 1} - 1}{\sqrt{\frac{4k}{k_c} + 1} + 1} \right)^{n_0} \right]. \quad (14)$$

We introduce the following characteristic scales of time and backtrack position,

$$n_c = \sqrt{\frac{4k}{k_c}}, \quad (15)$$

$$\tau_c = \frac{1}{k_c}. \quad (16)$$

The mean recovery time then simplifies to

$$\langle \tau_{\text{rec}} \rangle = \tau_c \left[1 - \left(\frac{\sqrt{n_c^2 + 1} - 1}{\sqrt{n_c^2 + 1} + 1} \right)^{n_0} \right]. \quad (17)$$

For initial positions that are not large compared to n_c , i.e. when $n_0 \leq n_c$, the mean recovery time given by

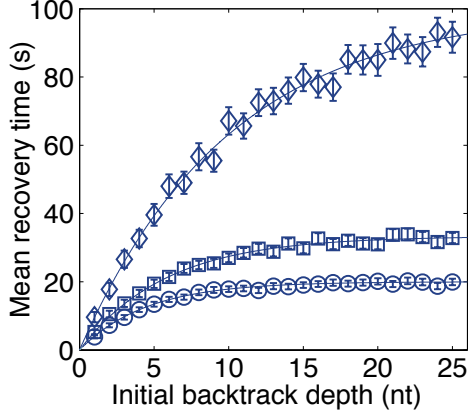


FIG. 5: **Mean recovery time as a function of the backtrack depth, discrete model.** Recovery time averaged over 1000 numerical simulations (symbols) of recovery from different initial backtrack depths. Diffusion rate was set to $k = 1/s$ in all cases and $k_c = 0.01/s$ (diamonds), $k_c = 0.03/s$ (squares) and $k_c = 0.05/s$ (circles). Error bars are standard errors of the mean with 90% statistical significance. The solid curves are obtained with the analytical expression given by Eq. (14) for the specific values of n_0 , k and k_c in this example.

Eq. (17) depends linearly on the initial backtrack depth

$$\langle \tau_{\text{rec}} \rangle = \tau_c \ln \left(\frac{\sqrt{n_c^2 + 1} + 1}{\sqrt{n_c^2 + 1} - 1} \right) n_0 + O(n_0^2) \quad (18)$$

When the initial position is much larger than the characteristic backtrack position n_c , i.e. when $n_0 \gg n_c$, the mean recovery time saturates to $\langle \tau_{\text{rec}} \rangle \rightarrow \tau_c$.

If the hopping and cleavage rates are equal ($k = k_c$), the mean recovery time given by Eq. (17) can be rewritten in terms of the Golden ratio $\varphi = (\sqrt{5} + 1)/2$ and the Golden ratio conjugate $\Phi = (\sqrt{5} - 1)/2$

$$\langle \tau_{\text{rec}} \rangle_{k=k_c} = \tau_c \left[1 - \left(\frac{\Phi}{\varphi} \right)^{n_0} \right] \quad (19)$$

The duration of a transcriptional pause, or equivalently the mean recovery time from $n_0 = 1$ is, for the case where $k = k_c$ equal to $\langle \tau_{\text{rec}} \rangle_{k=k_c, n=1} = \tau_c / \varphi$.

Figure 5 shows the analytical expression of the mean recovery time in the discrete model (14) compared to the average recovery time obtained from numerical simulations. The mean recovery time increases with increasing initial backtrack depth and saturates at $1/k_c$ for deep initial backtracks. The saturation of the mean recovery time at large n_0 was observed experimentally for Pol II recovery assisted with TFIIS [7].

In the absence of cleavage, the mean recovery time is not bounded, yielding $\langle \tau_{\text{rec}} \rangle = \infty$. Alternative statistics should therefore be considered to characterize the recovery in the absence of cleavage, such as the mode or the median recovery times.

III. CONTINUOUS MODEL: DIFFUSION PROCESS WITH CLEAVAGE

To address which features of the backtrack recovery process depend on the details of the 1D lattice of the DNA template, we now consider a continuous-space model where the motion of the polymerase is described by a diffusion process with a stochastic resetting [23] to the elongation state due to RNA cleavage. Such model can be envisioned as the continuous limit of the model in Fig. 3.

We consider that the position of the polymerase, x , is a continuous random variable. We define $\rho(x, t|x_0, 0)dx$ as the probability of a polymerase to be in the interval $[x, x + dx]$ at time t , given that it was at x_0 at time 0. In this continuous-space description the probability density $\rho(x, t|x_0, 0)$ evolves in time according to a Fokker-Planck equation with a diffusion term and a sink term,

$$\frac{\partial \rho(x, t|x_0, 0)}{\partial t} = D \frac{\partial^2 \rho(x, t|x_0, 0)}{\partial x^2} - k_c \rho(x, t|x_0, 0) \quad (20)$$

where we assume $x > 0$. Equation (20) results from taking the continuous limit in Eq. (2) and defining $x = an$, $x_0 = an_0$ and the diffusion coefficient $D = a^2k$, with $a = 0.34$ nm the distance between two nucleotides. The solution of the Fokker-Planck equation (20) with initial condition $\rho(x, 0|x_0, 0) = \delta(x - x_0)$ and the absorbing boundary condition $\rho(0, t|x_0, 0) = 0$ for $x > 0$ is given by [36]:

$$\rho(x, t|x_0, 0) = \frac{e^{-k_c t}}{\sqrt{4\pi Dt}} \left[e^{-(x-x_0)^2/4Dt} - e^{-(x+x_0)^2/4Dt} \right] \quad (21)$$

The recovery time probability density is given by the probability density flux to $x = 0$ due to diffusion, plus the probability flux due to cleavage,

$$\Phi(\tau_{\text{rec}}; x_0) = \Phi_{\text{diff}}(\tau_{\text{rec}}; x_0) + \Phi_c(\tau_{\text{rec}}; x_0) \quad (22)$$

$$= D \frac{\partial \rho(x, \tau_{\text{rec}}|x_0, 0)}{\partial x} \Big|_{x=0} + k_c S(\tau_{\text{rec}}; x_0) \quad (23)$$

where $\Phi_{\text{diff}}(\tau_{\text{rec}}; x_0)d\tau_{\text{rec}}$ is the probability to recover by diffusion in the time interval $[\tau_{\text{rec}}, \tau_{\text{rec}} + d\tau_{\text{rec}}]$ and $\Phi_c(\tau_{\text{rec}}; x_0)d\tau_{\text{rec}}$ is the probability to recover by cleavage in the time interval $[\tau_{\text{rec}}, \tau_{\text{rec}} + d\tau_{\text{rec}}]$. The probability density flux across $x = 0$ due to diffusion equals to

$$\Phi_{\text{diff}}(\tau_{\text{rec}}; x_0) = D \frac{\partial \rho(x, \tau_{\text{rec}}|x_0, 0)}{\partial x} \Big|_{x=0} \quad (24)$$

$$= e^{-k_c \tau_{\text{rec}}} \frac{x_0}{\sqrt{4\pi D \tau_{\text{rec}}^3}} e^{-x_0^2/4D\tau_{\text{rec}}} \quad (25)$$

where the spatial derivative in Eq. (24) is a derivative from the right. The survival probability at time τ_{rec} can

be calculated by integrating the probability of the polymerase to be at time τ_{rec} in $x > 0$,

$$S(\tau_{\text{rec}}; x_0) = \int_0^\infty \rho(x, \tau_{\text{rec}} | x_0, 0) dx \quad (26)$$

$$= e^{-k_c \tau_{\text{rec}}} \text{erf} \left(\frac{x_0}{\sqrt{4D\tau_{\text{rec}}}} \right) \quad (27)$$

where erf is the error function. The probability density $R(\tau_{\text{rec}}; x_0)$ of recovery from an initial backtrack depth x_0 in a time τ_{rec} , (referred to as the recovery probability) is then given by

$$R(\tau_{\text{rec}}; x_0) = 1 - S(\tau_{\text{rec}}; x_0) = 1 - e^{-k_c \tau_{\text{rec}}} \text{erf} \left(\frac{x_0}{\sqrt{4D\tau_{\text{rec}}}} \right). \quad (28)$$

For the case $k_c = 0$, the recovery probability simplifies to

$$R(\tau_{\text{rec}}; x_0) = \text{erfc} \left(\frac{x_0}{\sqrt{4D\tau_{\text{rec}}}} \right) \quad (29)$$

where erfc is the complementary error function. From Eq. (27), we obtain the probability density flux through $x = 0$ via cleavage:

$$\Phi_c(\tau_{\text{rec}}; x_0) = k_c S(\tau_{\text{rec}}; x_0) = k_c e^{-k_c \tau_{\text{rec}}} \text{erf} \left(\frac{x_0}{\sqrt{4D\tau_{\text{rec}}}} \right). \quad (30)$$

We obtain an exact expression for the recovery time distribution in the continuous-space model by adding Eq. (25) to (30)

$$\begin{aligned} \Phi(\tau_{\text{rec}}; x_0) &= e^{-k_c \tau_{\text{rec}}} \frac{x_0}{\sqrt{4\pi D \tau_{\text{rec}}^3}} e^{-x_0^2/4D\tau_{\text{rec}}} \\ &+ k_c e^{-k_c \tau_{\text{rec}}} \text{erf} \left(\frac{x_0}{\sqrt{4D\tau_{\text{rec}}}} \right). \end{aligned} \quad (31)$$

We now write Eq. (31) scaling time with respect to $\tau_c = 1/k_c$ and the initial position with respect to $x_c = \sqrt{4D/k_c}$ similarly to Eqs. (15) and (16) in the hopping model. In units of a scaled time $t_{\text{rec}} = \tau_{\text{rec}}/\tau_c$ and a scaled initial position $x_0 = x_0/x_c$, we obtain a universal expression

$$\Phi(t_{\text{rec}}; x_0) = e^{-t_{\text{rec}}} \frac{x_0}{\sqrt{\pi t_{\text{rec}}^3}} e^{-x_0^2/t_{\text{rec}}} + e^{-t_{\text{rec}}} \text{erf} \left(\frac{x_0}{\sqrt{t_{\text{rec}}}} \right). \quad (32)$$

To test the validity of the analytical expression for the recovery time distribution (32), we perform numerical simulations of the continuous model (see Fig. 6). The following overdamped Langevin equation describes the evolution of the backtracked distance at time t in continuous space, denoted as $x(t)$, $dx(t)/dt = \xi(t)$ where $\xi(t)$ models a stochastic force that drives the polymerase forward or backward. The stochastic force is described by

a delta-correlated Gaussian white noise with zero mean $\langle \xi(t) \rangle = 0$ and an amplitude proportional to the diffusion coefficient, $\langle \xi(t)\xi(t') \rangle = 2D\delta(t-t')$. Cleavage events are modelled as a stochastic resetting process [23] whose probability to occur in a time t is exponential $P_{\text{cleav}}(t) = k_c e^{-k_c t}$. We implement numerical simulations of the Langevin equation using an Euler discrete-time numerical integration scheme with $\Delta t = 1$ ms, which is one order of magnitude smaller than any characteristic time of backtrack recovery given by the inverse of cleavage or diffusion rates [7]. The results shown in Fig. 6 validate the exact expression obtained for the recovery time distribution given by Eq. (31) both in the presence and in the absence of cleavage. The recovery time distributions obtained for the same initial backtrack distance $x_0 = 5$ have the same shape as those obtained in the discrete-space description [cf. Fig. 4].

A. Mean recovery time

In the continuous model, the mean recovery time can be obtained by calculating the mean value of the first-passage distribution [Eq. (31)]

$$\langle \tau_{\text{rec}} \rangle = \frac{1}{k_c} \left[1 - e^{-x_0/\sqrt{D/k_c}} \right] \quad (33)$$

or equivalently

$$\langle \tau_{\text{rec}} \rangle = \tau_c \left[1 - e^{-2x_0/x_c} \right] \quad (34)$$

Note that, the mean recovery time can be also calculated by a different route, using the backward Fokker-Planck equation together with the Laplace transform of the survival probability (see Appendix B).

Equations (33) and (34) show that the mean recovery time for deep initial backtracks ($x_0 \gg x_c$) saturates to τ_c . Our results indicate that recovery happens mostly by diffusion for shallow backtracks, where $x_0 \ll x_c$, and mostly by cleavage for deep backtracks, where $x_0 \gg x_c$. For shallow initial backtracks ($x_0 \leq x_c$), the mean recovery time scales linearly with x_0 ,

$$\frac{\langle \tau_{\text{rec}} \rangle}{\tau_c} = \frac{x_0}{x_c/2} + O(x_0^2) \quad (35)$$

similarly to the mean recovery time in the discrete hopping model [see Eq. (18)].

Taking the limit $n_c \gg 1$ in the expression for the mean recovery time in the discrete model [Eq. (17)], we obtain

$$\langle \tau_{\text{rec}} \rangle \simeq \tau_c \left[1 - e^{-2n_0/(\sqrt{n_c^2+1})} \right] \simeq \tau_c \left[1 - e^{-2x_0/x_c} \right] \quad (36)$$

where we have used $x_0 = an_0$ and $x_c = an_c$. Note that this expression is equal to the mean recovery time in the continuous model [Eq. (34)]. Hence, the mean recovery times in the discrete and continuous description coincide for large characteristic depth.

Continuous diffusion model

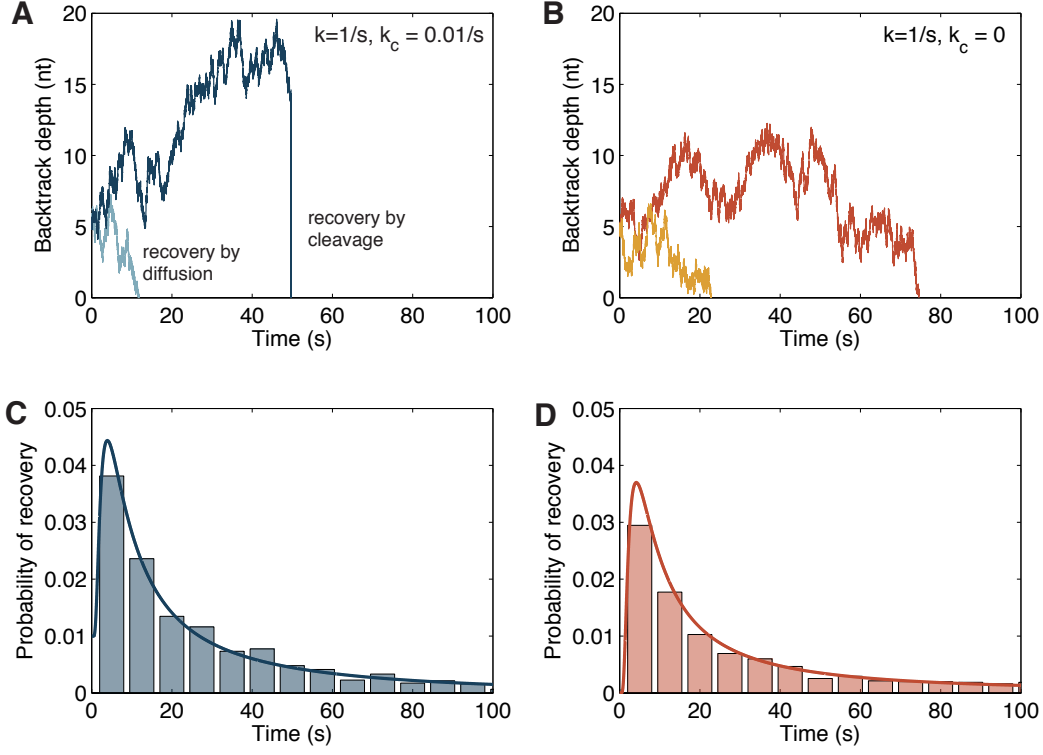


FIG. 6: **Stochastic trajectories of the continuous diffusion model and recovery time distributions.** **A)** Sample trajectories of the continuous-space model with diffusion and cleavage ($k = 1/s$, $k_c = 0.01/s$) simulated using Euler numerical scheme for a discrete-time Langevin equation. The light blue trajectory represents a polymerase that recovers by diffusion, and the dark blue trajectory a polymerase that recovers by cleavage. **B)** Sample trajectories for the continuous model with only diffusion, $k = 1/s$, $k_c = 0$ obtained with the same numerical integration scheme. **C)** Recovery time probability density for the case where $k = 1/s$ and $k_c = 0.01/s$. The bars are histograms obtained from 1000 numerical simulations and the curve is the exact expression given by Eq. (31). **D)** Recovery time probability density for the case where $k = 1/s$ and $k_c = 0$. The bars are histograms obtained from 1000 numerical simulations and the curve is the exact expression given by Eq. (31) with $k_c = 0$. In all simulations, the simulation time step was set to $\Delta t = 1$ ms and the initial distance to $x_0 = 5$ nt.

A comparison between the mean recovery time in the discrete and continuous model illustrates that both agree well for $n_c \geq 1$. In this regime, the polymerase typically performs a large number of jumps prior to recovery. Hence, the lattice does not impact on the mean recovery time, even for shallow initial backtracks.

Notably, the limit $n_c \geq 1$ is in agreement with the experimental data obtained from single molecule experiments with RNA polymerases II, where $k \sim 1 \text{ s}^{-1}$ and $k_c \sim (0.01 - 0.1) \text{ s}^{-1}$ [7, 10–12, 37], yielding $n_c > 1$. Therefore, for reported values of diffusion and cleavage rates, the diffusion approximation can be used without loss of generality, with the advantage of providing a simpler mathematical framework with respect to the hopping process.

IV. DISCUSSION

We have provided exact results on the statistics of the time needed for an RNA polymerase to recover from an arbitrary initial backtrack depth using discrete (hopping) and continuous-space (diffusion) stochastic descriptions. We have presented a road-map for the calculation of the first-passage time distribution for a continuous-time random walk with an absorbing state, which models RNA polymerase backtrack recovery with high fidelity. Both hopping and diffusion models provide similar recovery time distributions, with the majority of differences in the short recovery times and a complete overlap for long recovery times (see Table I for a summary of the main results).

We show that both discrete and continuous description can be used concurrently for backtrack recovery analysis for short and long backtracks when the characteristic distance $n_c = 2\sqrt{k/k_c}$ is greater than one. This corresponds to cases where the hopping rate k is larger than

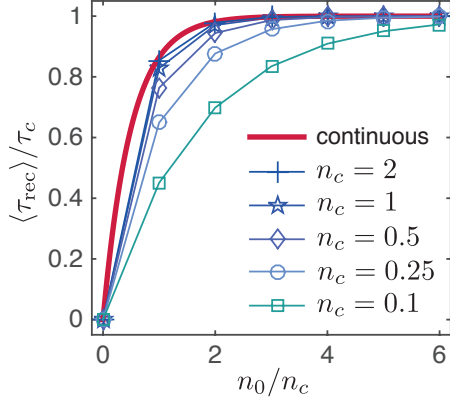


FIG. 7: **Scaled mean recovery time as a function of the scaled initial backtrack depth for discrete and continuous models.** Scaled mean recovery time $\langle \tau_{\text{rec}} \rangle / \tau_c$ as a function of the scaled initial backtrack depth n_0 / n_c for the discrete-space hopping model [Eq. (17), blue symbols] and as a function of x_0 / x_c for the continuous-space diffusion model [Eq. (34), magenta curve].

the cleavage rate k_c and is in good agreement with estimated rates of RNA polymerase backtracking [7, 11, 12].

Future work in the framework of stochastic resetting will have to be done to consider the case where polymerases can only cleave until a critical backtrack distance, as recently found in single-molecule experiments [7]. Single-molecule optical tweezers transcription experiments of RNA polymerase backtracking [6–8, 10–12, 22, 38] would allow to experimentally validate the stochastic models provided here and quantify the backtrack diffusion and cleavage rates of these enzymes.

V. ACKNOWLEDGMENTS

We are thankful to Shamik Gupta, Izaak Neri, Federico Vazquez, Masatoshi Nishikawa, Clélia de Mulatier and Francesc Font-Clos for helpful discussions. ER acknowledges financial support from Spanish Government, grants ENFASIS (FIS2011-22644) and TerMic (FIS2014-52486-R). ER and DST are thankful to Centro de Ciencias de Benasque Pedro Pascual (Benasque, Spain) for its hospitality. S.W.G. was supported by the EMBO Young Investigator Program, the Paul Ehrlich Foundation and grant no. 281903 from the European Research Council.

TABLE I: Summary of expressions for the probability distribution of the recovery time and the mean recovery time from a given initial backtrack depth in the hopping model with cleavage $\Phi(\tau_{\text{rec}}; n_0)$ and in the diffusion model with cleavage $\Phi(\tau_{\text{rec}}; x_0)$ with initial backtrack depths n_0 and x_0 respectively. Here, k is the hopping rate, D is the diffusion coefficient and $H(\tau_{\text{rec}}; n_0) = {}_2F_2(\{n_0, n_0 + 1/2\}; \{n_0 + 1, 2n_0 + 1\}; -4k\tau_{\text{rec}})$.

Discrete hopping model		
	Diffusion and cleavage ($k > 0$; $k_c > 0$)	Only diffusion ($k > 0$; $k_c = 0$)
$\Phi(\tau_{\text{rec}}; n_0)$	$e^{-(2k+k_c)\tau_{\text{rec}}} \frac{n_0}{\tau_{\text{rec}}} \frac{I_{n_0}(2k\tau_{\text{rec}})}{k_c} + k_c e^{-k_c\tau_{\text{rec}}} \left[1 - \frac{(k\tau_{\text{rec}})^{n_0} H(\tau_{\text{rec}}; n_0)}{n_0 \Gamma(n_0)} \right]$	$e^{-(2k+k_c)\tau_{\text{rec}}} \frac{n_0}{\tau_{\text{rec}}} \frac{I_{n_0}(2k\tau_{\text{rec}})}{\tau_{\text{rec}}}$
$\langle \tau_{\text{rec}} \rangle$	$\frac{1}{k_c} \left[1 - \left(\frac{\sqrt{(4k/k_c)+1}-1}{\sqrt{(4k/k_c)+1}+1} \right)^{n_0} \right]$	∞
Continuous diffusion model		
	Diffusion and cleavage ($D > 0$; $k_c > 0$)	Only diffusion ($D > 0$; $k_c = 0$)
$\Phi(\tau_{\text{rec}}; x_0)$	$e^{-k_c\tau_{\text{rec}}} \frac{x_0}{\sqrt{4\pi D\tau_{\text{rec}}^3}} e^{-x_0^2/4D\tau_{\text{rec}}} + k_c e^{-k_c\tau_{\text{rec}}} \text{erf} \left(\frac{x_0}{\sqrt{4D\tau_{\text{rec}}}} \right)$	$\frac{x_0}{\sqrt{4\pi D\tau_{\text{rec}}^3}} e^{-x_0^2/4D\tau_{\text{rec}}}$
$\langle \tau_{\text{rec}} \rangle$	$\frac{1}{k_c} \left[1 - e^{-x_0/\sqrt{D/k_c}} \right]$	∞

[1] Nudler, E., Mustaev, A., Goldfarb, A. & Lukhtanov, E. The RNA-DNA hybrid maintains the register of tran-

scription by preventing backtracking of RNA polymerase. *Cell* **89**, 33–41 (1997).

- [2] Komissarova, N. & Kashlev, M. Transcriptional arrest: *E. coli* RNA polymerase translocates backward, leaving the 3'-end of the RNA intact and extruded. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 1755–1760 (1997).
- [3] Kettenberger, H., Armache, K.-J. & Cramer, P. Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage. *Cell* **114**, 347–357 (2003).
- [4] Wang, D. *et al.* Structural basis of transcription: backtracked RNA polymerase II at 3.4 angstrom resolution. *Science* **324**, 1203–1206 (2009).
- [5] Cheung, A. C. & Cramer, P. Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature* **471**, 249–253 (2011).
- [6] Galburt, E. A. *et al.* Backtracking determines the force sensitivity of RNAP II in a factor-dependent manner. *Nature* **446**, 820–823 (2007).
- [7] Lisica, A. *et al.* Mechanisms of backtrack recovery by rna polymerases i and ii. *Proceedings of the National Academy of Sciences* **113**, 2946–2951 (2016).
- [8] Shaevitz, J. W., Abbondanzieri, E. A., Landick, R. & Block, S. M. Backtracking by single RNAPolymerase molecules observed at near-base-pair resolution. *Nature* **426**, 684–687 (2003).
- [9] Depken, M., Galburt, E. A. & Grill, S. W. The origin of short transcriptional pauses. *Biophys. J.* **96**, 2189–2193 (2009).
- [10] Hodges, C., Bintu, L., Lubkowska, L., Kashlev, M. & Bustamante, C. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science* **325**, 626–628 (2009).
- [11] Dangkulwanich, M. *et al.* Complete dissection of transcription elongation reveals slow translocation of RNA polymerase II in a linear ratchet mechanism. *eLife* **2** (2013).
- [12] Ishibashi, T. *et al.* Transcription factors IIS and IIF enhance transcription efficiency by differentially modifying RNA polymerase pausing dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 3419–3424 (2014).
- [13] Kuhn, C.-D. *et al.* Functional architecture of RNA polymerase I. *Cell* **131**, 1260–1272 (2007).
- [14] Walmaecq, C. *et al.* Rpb9 subunit controls transcription fidelity by delaying NTP sequestration in RNA polymerase II. *J. Biol. Chem.* **284**, 19601–19612 (2009).
- [15] Chédin, S., Riva, M., Schultz, P., Sentenac, A. & Carles, C. The RNA cleavage activity of RNA polymerase III is mediated by an essential TFIIS-like subunit and is important for transcription termination. *Genes Dev.* **12**, 3857–3871 (1998).
- [16] Izban, M. G. & Luse, D. S. The RNA polymerase II ternary complex cleaves the nascent transcript in a 3'–5' direction in the presence of elongation factor SII. *Genes Dev.* **6**, 1342–1356 (1992).
- [17] Fish, R. N. & Kane, C. M. Promoting elongation with transcript cleavage stimulatory factors. *Biochim. Biophys. Acta* **1577**, 287–307 (2002).
- [18] Ruan, W., Lehmann, E., Thomm, M., Kostrewa, D. & Cramer, P. Evolution of two modes of intrinsic RNA polymerase transcript cleavage. *J. Biol. Chem.* **286**, 18701–18707 (2011).
- [19] Larson, M. H. *et al.* Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 6555–6560 (2012).
- [20] Kloppe, A. V., Bois, J. S. & Grill, S. W. Influence of secondary structure on recovery from pauses during early stages of RNA transcription. *Physical Review E* **81**, 1–4 (2010).
- [21] Sahoo, M. & Klumpp, S. Backtracking dynamics of RNA polymerase: pausing and error correction. *Journal of Physics: Condensed Matter* **25**, 374104 (2013).
- [22] Schweikhard, V. *et al.* Transcription factors TFIIF and TFIIS promote transcript elongation by RNA polymerase II by synergistic and independent mechanisms. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 6642–6647 (2014).
- [23] Evans, M. R. & Majumdar, S. N. Diffusion with stochastic resetting. *Phys. Rev. Lett.* **106**, 160601 (2011).
- [24] Evans, M. R. & Majumdar, S. N. Diffusion with optimal resetting. *Journal of Physics A: Mathematical and Theoretical* **44**, 435001 (2011).
- [25] Gupta, S., Majumdar, S. N. & Schehr, G. Fluctuating interfaces subject to stochastic resetting. *Phys. Rev. Lett.* **112**, 220601 (2014).
- [26] Durang, X., Henkel, M. & Park, H. The statistical mechanics of the coagulation–diffusion process with a stochastic reset. *Journal of Physics A: Mathematical and Theoretical* **47**, 045002 (2014).
- [27] Nagar, A. & Gupta, S. Diffusion in presence of stochastic resetting at power-law times. *arXiv preprint arXiv:1512.02092* (2015).
- [28] Pal, A. Diffusion in a potential landscape with stochastic resetting. *Physical Review E* **91**, 012113 (2015).
- [29] Rotbart, T., Reuveni, S. & Urbakh, M. Michaelis-Menten reaction scheme as a unified approach towards the optimal restart problem. *Physical Review E* **92**, 060101 (2015).
- [30] Reuveni, S. Optimal stochastic restart renders fluctuations in first passage times universal. *arXiv preprint arXiv:1512.01600* (2015).
- [31] Jahnel, M., Behrndt, M., Jannasch, A., Schäffer, E. & Grill, S. W. Measuring the complete force field of an optical trap. *Optics letters* **36**, 1260–1262 (2011).
- [32] Depken, M., Parrondo, J. M. R. & Grill, S. W. Intermittent transcription dynamics for the rapid production of long transcripts of high fidelity. *Cell Rep.* **5**, 521–530 (2013).
- [33] Gardiner, C. W. *Stochastic methods* (Springer-Verlag, Berlin, Germany, 2009).
- [34] Abramowitz, M. & Stegun, I. A. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. 55 (Courier Corporation, 1964).
- [35] Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.* **22**, 403–434 (1976).
- [36] Redner, S. *A guide to first-passage processes* (Cambridge University Press, 2001).
- [37] Zamft, B., Bintu, L., Ishibashi, T. & Bustamante, C. Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8948–8953 (2012).
- [38] Abbondanzieri, E. A., Greenleaf, W. J., Shaevitz, J. W., Landick, R. & Block, S. M. Direct observation of base-pair stepping by RNA polymerase. *Nature Cell Biology* **438**, 460–465 (2005).

VI. APPENDIX A: EXACT SOLUTION OF THE HOPPING MODEL WITH DIFFUSION

Equations (1-2) can be rewritten as

$$\frac{d}{dt}P(t) = AP(t) \quad . \quad (37)$$

where $P(t) = [p_1(t) p_2(t) \dots]^\top$ is a column vector including the state probabilities at time t and A is a tridiagonal symmetric Toeplitz matrix [?] of the form

$$A = \begin{bmatrix} -(2k+k_c) & k & 0 & 0 & \dots \\ k & -(2k+k_c) & k & 0 & \dots \\ 0 & k & -(2k+k_c) & k & 0 \dots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix}.$$

The solution of Eq. (37) with initial condition $P(0) = [0, 0, \dots, 0, 1, 0, \dots]^\top$, with $p_{n_0}(0) = 1$ and $p_n(0) = 0$ for $n \neq n_0$, is given by [?]

$$P(t) = P(0)e^{At} \quad . \quad (38)$$

We now decompose A in the following form

$$A = Q D Q^{-1} \quad . \quad (39)$$

where D is a diagonal matrix containing the eigenvalues of A and Q is a matrix with the eigenvectors of A in columns. Note that since A is symmetric, $Q^{-1} = Q^\top$. To obtain the eigenvalues of A we first assume that the matrix is of finite size $N \times N$ and then take the limit $N \rightarrow \infty$. For N finite, the matrix elements of the matrices D and Q are given by [?]

$$D_{ii} = -(2k+k_c) + 2k \cos\left(\frac{i\pi}{N+1}\right) \quad , \quad (40)$$

$$Q_{ij} = \sqrt{\frac{2}{N+1}} \sin\left(\frac{ij\pi}{N+1}\right) \quad . \quad (41)$$

The term $\sqrt{\frac{2}{N+1}}$ that appears in Eq. (41) is the normalization constant. As a result, Eq. (38) can be rewritten as

$$P(t) = Q e^{Dt} Q^\top P(0) \quad . \quad (42)$$

where e^{Dt} is a diagonal matrix with elements $e^{D_{ii}t}$ ($i = 1, \dots, N$) in the diagonal. After some algebra, we obtain the following expression for the n -th element of the vector $P(t)$

$$p_n(t) = \sum_{m=1}^N \sin\left(\frac{kn\pi}{N+1}\right) e^{[-(2k+k_c) + 2k \cos(\frac{m\pi}{N+1})]t} \times \frac{2}{N+1} \sin\left(\frac{n_0 k \pi}{N+1}\right) \quad . \quad (43)$$

We now take the asymptotic limit $N \rightarrow \infty$. In this limit, $k/N \rightarrow x$ where x is a continuous variable and the

sum $\sum_{k=1}^N \frac{1}{N} \rightarrow \int_0^1 dx$. Using these approximations, we obtain

$$\begin{aligned} p_n(t) &= \frac{2e^{-(2k+k_c)t}}{\pi} \int_0^\pi e^{2kt \cos x} \sin(n_0 x) \sin(nx) dx \\ &= \frac{e^{-(2k+k_c)t}}{\pi} \left\{ \int_0^\pi e^{2kt \cos(x)} \cos[(n_0 - n)x] dx \right. \\ &\quad \left. - \int_0^\pi e^{2kt \cos(x)} \cos[(n_0 + n)x] dx \right\} \\ &= e^{-(2k+k_c)t} [I_{n_0-n}(2kt) - I_{n_0+n}(2kt)] \quad . \quad (44) \end{aligned}$$

Here we have used the property

$$\sin(ax) \sin(bx) = \frac{1}{2} \{ \cos[(a-b)x] - \cos[(a+b)x] \},$$

and the definition of the modified Bessel function of the first kind [34]

$$I_m(z) = \frac{1}{\pi} \int_0^\pi e^{z \cos x} \cos(mx) dx \quad . \quad (45)$$

From Eq. (44) we can obtain the probability to be at state $n = 1$ at time t ,

$$p_1(t) = e^{-(2k+k_c)t} [I_{n_0-1}(2kt) - I_{n_0+1}(2kt)] \quad . \quad (46)$$

Using the property $I_{m-1}(z) - I_{m+1}(z) = \frac{2m}{z} I_m(z)$ [34] in Eq. (46) we obtain

$$p_1(t) = e^{-(2k+k_c)t} \frac{n_0 I_{n_0}(2kt)}{kt} \quad , \quad (47)$$

which equals to Eq. (7).

VII. APPENDIX B: CALCULATION OF THE MEAN RECOVERY TIME FROM THE BACKWARD FOKKER-PLANCK EQUATION

The Backward Fokker-Planck equation corresponding to Eq. (20) in the continuous-space model reads

$$\frac{\partial \rho(x, t|x_0, 0)}{\partial t} = D \frac{\partial^2 \rho(x, t|x_0, 0)}{\partial x_0^2} - k_c \rho(x, t|x_0, 0) \quad . \quad (48)$$

Integrating Eq. (49) with respect to x from $x = 0$ to $x = \infty$ we obtain the following equation for the survival probability:

$$\frac{\partial S(t; x_0)}{\partial t} = D \frac{\partial^2 S(t; x_0)}{\partial x_0^2} - k_c S(t; x_0) \quad . \quad (49)$$

Taking the Laplace transform, Eq. (49) yields

$$q\mathbb{S}(q; x_0) - 1 = D \frac{\partial^2 \mathbb{S}(q; x_0)}{\partial x_0^2} - k_c \mathbb{S}(q; x_0) \quad , \quad (50)$$

where $\mathbb{S}(q; x_0) = \int_0^\infty dt e^{-qt} S(t; x_0)$ is the Laplace transform of the survival probability and we have used $S(0; x_0) = 1$. The solution of Eq. (50) is given by

$$\mathbb{S}(q; x_0) = \frac{1}{k_c + q} \left[1 - e^{-x_0 \sqrt{(k_c + q)/D}} \right] \quad . \quad (51)$$

From Eq. (51) one can find all the moments of the recovery time distribution. In particular, the mean recovery time:

$$\langle \tau_{\text{rec}} \rangle = \mathbb{S}_n(0) = \frac{1}{k_c} \left[1 - e^{-x_0 \sqrt{k_c/D}} \right] \quad , \quad (52)$$

which coincides with Eq. (33).